EOHA

DESIGN SPECIFICATION DOCUMENT

MEDICAL CONSORTIUM
OF AFRICA (PTY) LTD

Prepared By :

Simamnkele Mlisana
Msesenyane Makhongela

Medical Consortium of Africa
University of Pretoria
TuksNovation, 14th Floor
Humanities Building.

DESIGN SPECIFICATION

CLIENT OVERVIEW

The Medical Consortium of Africa is a health-focused organization committed to addressing some of the most pressing public health challenges in South Africa and across the continent. Its core aim is to improve public health outcomes using technology, data, and proactive disease surveillance.

PROJECT STAKEHOLDERS

Partner	Roles
TuksNovation Business Incubation Programme	Provides business incubation support, including mentorship, venture development guidance, and access to entrepreneurial resources to help bring the project to market.
NEOFrontiers Enterprise Innovation Support Fund (EISF)	Provides financial and strategic support for innovation, specifically focused on advancing Earth Observation (EO)-driven solutions.
South African National Space Agency (SANSA)	Contributes expertise and access to Earth Observation (EO) satellite data and supports the technical application of space-based technologies.
National Research Foundation (NRF)	Supports research development, funding, and alignment with national research priorities.

BUSINESS OPPORTUNITY

Public health systems are gradually shifting from reactive treatment to prevention. Improved disease surveillance particularly for malaria can help health authorities identify at-risk populations, anticipate outbreaks, and respond more effectively.

There is also a growing need for better monitoring of non-communicable diseases (NCDs). Strong surveillance systems allow governments to design more targeted interventions and evaluate their effectiveness over time.

The project will be introduced through academic platforms, data science forums, and partnerships with health organisations. A free trial version of the platform will be used to demonstrate its capabilities. Collaborations with organisations such as Target Malaria, as well as engagement with major global health funders, will support expansion into high-risk regions.

Target market

National Governments & Ministries of Health

Intergovernmental & Transnational Health Agencies

Global Philanthropies & Non-Profit Consortiums

Academic & Research Institutions

PROBLEM STATEMENT

- Malaria: Sub-Saharan Africa accounts for 94% of global cases and 95% of deaths.

- NCDs: Responsible for 37% of deaths in Sub-Saharan Africa, and a massive 51% in South Africa.

Africa faces a dual burden of disease: malaria and non-communicable diseases. According to the World Health Organization, NCD-related deaths in Sub-Saharan Africa increased from 24% in the early 2000s to 37% in 2019. At the same time, the region accounts for 94% of global malaria cases and 95% of malaria-related deaths.

A major contributing factor is the lack of effective health surveillance systems. Limited resources make it difficult for governments to plan, monitor, and respond to disease trends in a timely and data-driven way.

PROPOSED SYSTEM AND BENEFITS

The Earth Observational Health Analytics (EOHA) platform is designed to address these challenges by using satellite data and advanced analytics to monitor and predict disease patterns.

01

Social Impact

Led by a diverse, youth-driven team, including strong representation of women in STEM, the project contributes to transformation in the tech and health sectors.

02

Economic Impact

It will create opportunities for skilled youth in data analytics and Earth Observation, supporting job creation and SME development.

03

Academic Impact

In partnership with the University of Pretoria, the platform will support postgraduate research in fields such as remote sensing and public health.

SYSTEM FUNCTIONALITY

The EOHA platform integrates environmental and health data to detect and predict disease patterns, particularly malaria outbreaks.

Data Ingestion & Integration

- **EO Data Processing Pipeline:**
  Ingests and processes satellite data for malaria such as surface temperature, soil moisture, water bodies, and habitat type, as well as NCD-related environmental factors such as air pollution, UV radiation, and urban growth.
- **Ground Data Integration:**
  Combines epidemiological data such as case reports and patient movement with entomological data including mosquito surveillance and species mapping.
- **EO Data Cubes:**
  Structures large, multi-dimensional datasets for efficient querying and analysis.

- **Automated Pipelines:**
  Enables continuous data ingestion, cleaning, and integration with minimal manual intervention.

Artificial Intelligence & Analytics

- **Predictive Modelling:**
  Forecasts malaria outbreaks and models NCD trends using environmental risk factors.
- **Machine Learning:**
  Identifies patterns in integrated datasets and improves prediction accuracy over time.
- **Geospatial AI:**
  Analyses spatial patterns to detect high-risk areas and vulnerable populations.

Visualisation & User Interface

- **Risk Zone Mapping:**
  Displays current and predicted disease risk areas using GIS-based maps.
- **Digital Twin Environment:**
  Represents real-world environmental and public health conditions for monitoring and simulation.
- **AR/VR Visualisation:**
  Provides immersive environments for exploring data and model outputs.
- **Forecasting Dashboard:**
  Delivers near real-time insights through an interactive interface for decision-making.

FEASIBILITY

Technical Feasibility: The system converts complex satellite data into usable insights for public health decision-making. Development is structured into six milestones, from initial planning to full deployment with AI integration.

Team Feasibility: The team consists of young professionals with advanced qualifications in health and remote sensing. The group is predominantly female-led and brings both technical expertise and research experience.

NON-FUNCTIONAL REQUIREMENTS

Requirement	Explanation	Details
Functionality	The system must provide accurate and relevant outputs based on integrated environmental and health data.	• Secure login with role-based access • Accurate risk maps and disease analytics • Integration of EO, environmental, and health data • Data validation for consistency and accuracy
Usability	The platform must be intuitive and accessible to both technical and non-technical users	• Simple, dashboard-based interface • Clear visualisations • Minimal training required • Accessible via standard web platforms
Reliability	Given its role as an early warning system, the platform must be dependable	• High availability with minimal downtime • Handles incomplete or delayed data • Consistent and stable outputs • Backup and recovery in place
Performance	The system must handle large-scale data efficiently and operate close to real-time	• Processes large datasets efficiently • Near real-time data updates • Fast dashboard response times
Supportability	The system must be maintainable, extensible, and easy to support over time	• Modular and extensible design • Easy maintenance and updates • Well-documented system • Compatible with evolving data sources

PROJECT TIMELINE

01	Stakeholder & Partner Alignment	04	Rigorous Data, Sync, & Functionality Testing
02	Data Strategy & Infrastructure Setup	05	AI Integration & Advanced Analytics Implementation
03	Data Deployment & Prototype Ingestion	06	Final Improvements & Platform Launch

USE CASES AND MODELS

SYSTEM DESCRIPTION

The system encompasses the entire process of ingesting environmental data, tracking clinical health records, running predictive machine learning models, and delivering actionable disease forecasts to health officials within the Earth Observational Health Analytics (EOHA) platform. It includes the automated satellite data pipeline, the AI forecasting engine, the interactive spatial dashboard, and the proactive risk alerting functionalities.

System Components

01

Data Ingestion

Automates the retrieval and mapping of environmental satellite data and clinical ground truth observations to specific spatial zones.

02

Predictive ML

Runs the core machine learning algorithms against historical and environmental data to generate future risk scores and outbreak forecasts for Malaria and NCDs.

03

Geo Dashboard

Provides the interactive, map-based interface where users can view color-coded risk zones, historical health data, and AI predictions.

04

Alert System

Monitors new AI forecasts against predefined risk thresholds and automatically dispatches targeted warnings to authorized health officials based on their jurisdiction.

OVERVIEW OF ACTORS AND THEIR ROLES

The system manages spatial data ingestion, predictive risk modeling, localized observations, and user coordination through its automated pipelines, machine learning engine, geospatial dashboard, and alerting tools.

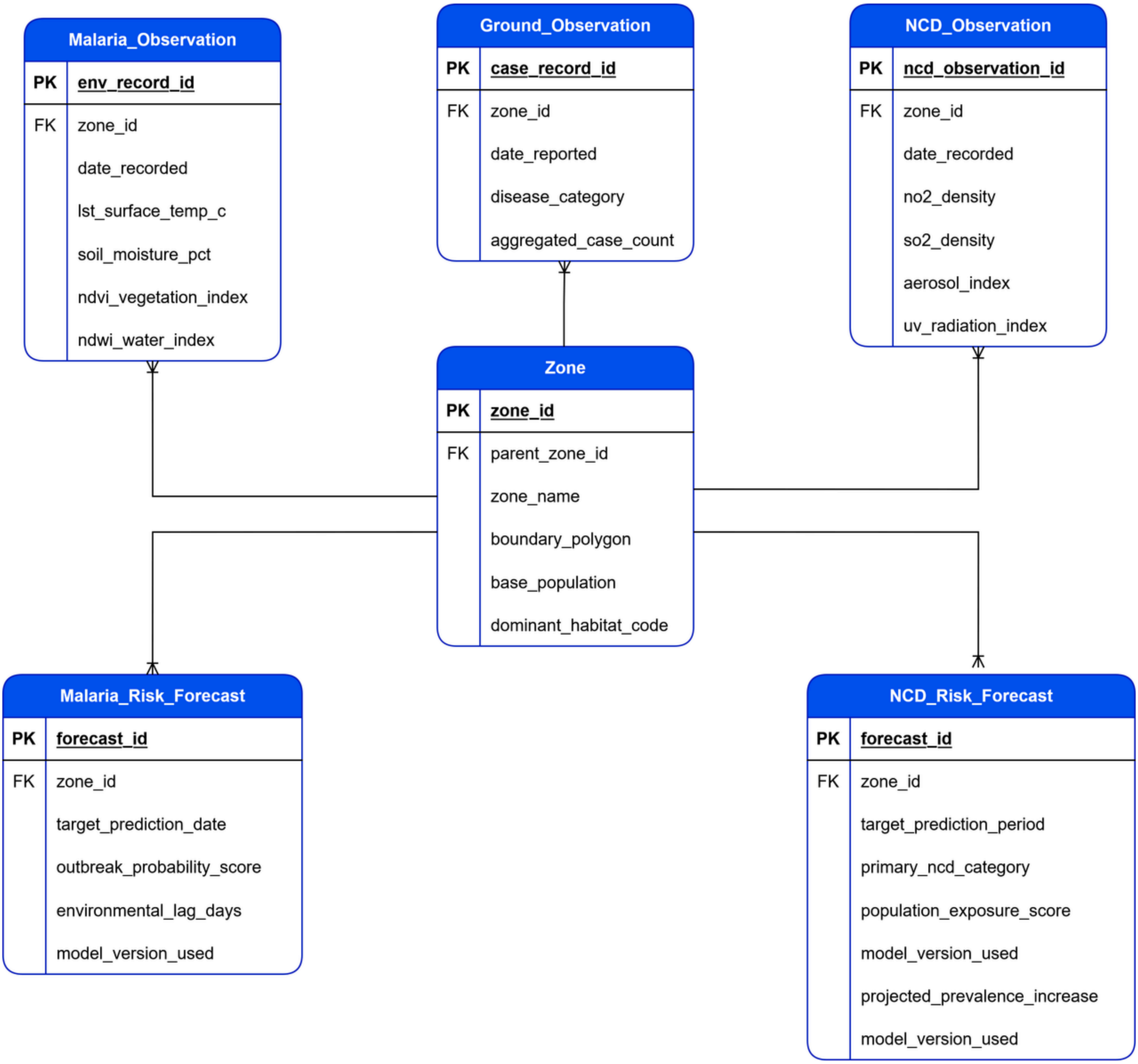
The following section defines the primary human and automated actors that interact with the platform. Outlining these roles provides a clear understanding of who engages with the system and their core responsibilities from data collection and administration to monitoring alerts and ensuring system integrity.

Actor	Roles
End User	• Browse interactive risk maps. • View predictive AI forecasts. • Track and monitor active alerts. • Acknowledge and update alert status
Data Contributor	• Log localized ground-truth data. • Report periodic observation counts. • Update records for specific geographic zones.
Administrator	• Oversee system access and overall operation. • Manage user roles and permissions. • Configure and update spatial zone boundaries. • Monitor system usage and performance.
System	• Facilitate seamless interaction between all users. • Ingest environmental data from external sources. • Execute predictive machine learning algorithms. • Render geospatial mapping data for dashboards. • Generate and dispatch automated risk alerts. • Ensure data integrity, security, and smooth functioning of all system components.

ENTITY RELATIONSHIP DIAGRAM

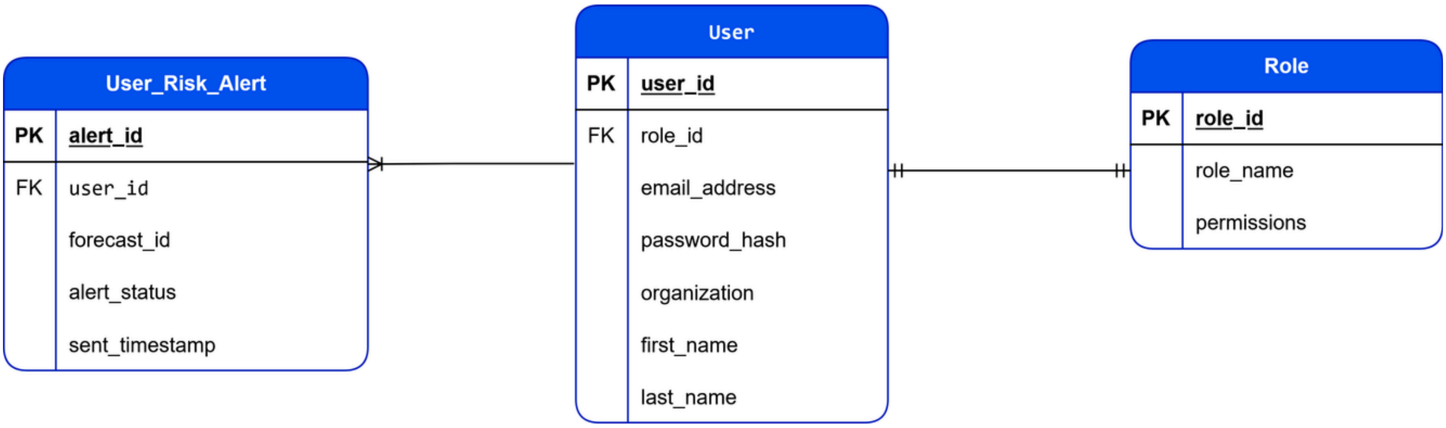
Observation Prediction ERD

This illustrates the logical database structure of the prediction system. As the central data repository, it maps out the core entities, such as spatial zones, environmental observations, and predictive forecasts, along with their relationships. This structural blueprint ensures strict data integrity and supports the back-end APIs in securely serving information to the front-end client.



User Management ERD

This illustrates the logical data structure for the user management and notification components of the platform. It maps out the core entities responsible for access control and client interactions, specifically focusing on user profiles, system roles, and personalized risk alerts. This structural blueprint defines how user identities are organized, authenticated, and securely linked to targeted communications.



USE CASE SET

The following section outlines the primary use cases for the platform. A use case set defines the functional requirements of the system by detailing the specific interactions between the defined actors and the application. By mapping out these practical scenarios, we establish a clear understanding of how users navigate the system to achieve core objectives such as data management, spatial risk monitoring, and alert acknowledgment.

Observation Prediction Use Case Set

Actor	Roles	
Zone	• Create Zone (System/Admin) • Read Zone (System/User) • Update Zone (System/Admin)	• Read Zone(includes) *(for parent_zone_id)* • Read Zone(includes)
Malaria_Observation	• Create Malaria_Observation (System) • Read Malaria_Observation (System/User)	• Read Zone(includes) • Read Zone(includes)
NCD_Observation	• Create NCD_Observation (System) • Read NCD_Observation (System/User)	• Read Zone(includes) • Read Zone(includes)
Ground_Observation	• Create Ground_Observation (System/User) • Read Ground_Observation (System/User)	• Read Zone(includes) • Read Zone(includes)
Malaria_Risk_Forecast	• Create Malaria_Risk_Forecast (System/ML) • Read Malaria_Risk_Forecast (User/System)	• Read Zone(includes), Read Malaria_Observation(includes), Read Ground_Observation(includes) • Read Zone(includes)
NCD_Risk_Forecast	• Create NCD_Risk_Forecast (System/ML) • Read NCD_Risk_Forecast (User/System)	• Read Zone(includes), Read NCD_Observation(includes), Read Ground_Observation(includes) • Read Zone(includes)

User Management Use Case Set

Actor	Roles	
Role	• Create Role (Admin) • Read Role (System/Admin) • Update Role (Admin)	• Read Role(includes)
User	• Create User (Admin) • Read User (System/User) • Update User (User/Admin)	• Read Role(includes) • Read Role(includes) • Read User(includes), Read Role(includes)
User_Risk_Alert	• Create Alert (System) • Read Alert (User/System) • Update Alert Status (User/System)	• Read User(includes), Read Forecast(includes) • Read User(includes) • Read Alert(includes)

USE CASE DESCRIPTIONS

The following section provides detailed, step-by-step descriptions for each core use case within the platform. These descriptions break down the specific triggering events, involved actors, preconditions, and the sequential flow of activities required to complete a process. By mapping out the main success scenarios and post-conditions, this section serves as a comprehensive guide to understanding exactly how user interactions and automated background tasks are executed to achieve the system's objectives.

Create Malaria_Observation

This use case involves the automated system ingesting new environmental data from satellites (like Google Earth Engine) to track conditions that lead to malaria outbreaks. The main success scenario includes the system pulling satellite imagery, mapping it to specific spatial zones, and saving the temperature and moisture data.

Use Case Name	Create Malaria_Observation
Triggering Event	A sufficient batch of new Malaria_Observation data is successfully ingested into the database.
Actor(s)	**ML Pipeline: The Python-based AI model that processes data and outputs a forecast.**
Related Use Cases	• Read Zone • Read Malaria_Observation • Read Ground_Observation • Create User_Risk_Alert
Pre-condition(s)	• Sufficient historical training data must exist in the database. {ObservationCount > Minimum Threshold} • The ML model weights must be loaded and active.
Flow of Activities	1. **External API (GEE)**: Generates the latest 8-day satellite imagery for the continent. 2. System: Triggers a request to pull data for specific coordinates. 3. **External API (GEE)**: Returns a JSON payload containing surface temperature, soil moisture, and water indexes. • System: Parses the JSON and maps the coordinates to the corresponding zone_id (READ ZONE). 4. **Database**: Validates the zone_id. 5. **System**: Formats the payload for database insertion. 6. **System**: Executes the INSERT command to save the new observations.
Post-condition	A new environmental record must have been created in the **MALARIA_OBSERVATION** table using attributes [env_record_id, zone_id, date_recorded, lst_surface_temp_c, soil_moisture_pct].

Create Malaria_Risk_Forecast

This use case involves the Machine Learning pipeline running its algorithms to predict future disease outbreaks. The main success scenario involves the AI pulling historical weather and health data, running the XGBoost model, and saving a probability score for a specific zone.

Use Case Name	Create Malaria_Risk_Forecast
Triggering Event	The scheduled integration service triggers (e.g., every 8 days when a new satellite pass completes).
Actor(s)	• System Integration Service: Initiates the request to the external satellite API. • External API (GEE): Provides the raw environmental data.
Related Use Cases	• Read User_Risk_Alert • Read Malaria_Risk_Forecast
Pre-condition(s)	• The user must be authenticated and authorized to view the specific zone_id. • An alert must exist in the database with an alert_status of 'Sent' or 'Active'.
Flow of Activities	1. ML Pipeline: Initiates the weekly prediction sequence. 2. System: Queries the required historical features. • Database: Retrieves lst_surface_temp_c and soil_moisture_pct (READ MALARIA_OBSERVATION). • Database: Retrieves past aggregated_case_count (READ GROUND_OBSERVATION). 3. ML Pipeline: Ingests the queried features, applies the environmental lag days, and runs the XGBoost algorithm. 4. ML Pipeline: Outputs an outbreak_probability_score and expected_case_surge for each zone_id. • Database: Validates the zone_id. 5. System: Prepares the prediction data for storage. 6. System: Sends the finalized data to the database. • Database: A new forecast record is added to the MALARIA_RISK_FORECAST table. • System: If the outbreak_probability_score > 0.80, it triggers the {Create User_Risk_Alert} use case.
Post-condition	A new forecast record is created in the **MALARIA_RISK_FORECAST** table using attributes [forecast_id, zone_id, target_prediction_date, outbreak_probability_score].

Update Alert Status

This use case involves a human user interacting with the early-warning system. The main success scenario includes the user seeing a high-risk alert on their dashboard, reviewing the map, and acknowledging the warning to coordinate an on-the-ground response.

Use Case Name	Update Alert Status
Triggering Event	The user clicks the "Acknowledge" or "Resolve" button on an active risk alert inside the Leaflet dashboard.
Actor(s)	• User: Reviews the data and updates the alert.
Related Use Cases	• Read User_Risk_Alert • Create Malaria_Risk_Forecast
Pre-condition(s)	• An alert must exist in the database with an alert_status of 'Sent' or 'Active'.
Flow of Activities	1. **User:** Logs into the dashboard and navigates to the "Active Alerts" panel. • **System:** Fetches active alerts for the user's assigned zones. • **Database:** Returns rows from the USER_RISK_ALERT table where the status is 'Active'. 2. **User:** Clicks on a specific alert to view its details. • **System:** Fetches the corresponding ML prediction details. • **Database:** Returns prediction data from the MALARIA_RISK_FORECAST table. 3. **User:** Reviews the outbreak_probability_score and the map boundaries. 4. **User:** Clicks the 'Acknowledge Alert' button to confirm they are taking action. 5. **System:** Receives the state change request. 6. **System:** Executes the update. • **Database:** The alert_status in the USER_RISK_ALERT table is updated to 'Acknowledged'.
Post-condition	• The alert record in the **USER_RISK_ALERT** table is updated. • The alert is removed from the user's "Active" UI view and moved to "Historical/Resolved".

Read Malaria_Risk_Forecast

This use case involves a User interacting with the geospatial map to view AI predictions. The main success scenario includes the system fetching spatial boundaries and overlaying them with color-coded risk forecasts generated by the ML pipeline.

Use Case Name	Malaria_Risk_Forecast
Triggering Event	The user navigates to the "Risk Map" dashboard view.
Actor(s)	• User: Views the interactive risk maps.
Related Use Cases	• Create Malaria_Risk_Forecast
Pre-condition(s)	• Forecasts exist for the current date range.
Flow of Activities	1. **User:** Clicks on the "Risk Map" tab in the frontend interface. 2. **System:** Triggers an API request to fetch map boundaries and active forecasts. • **Database:** Retrieves polygon boundaries from the ZONE table *(Note: added from your earlier use case details for completeness!)*. • **Database:** Retrieves the latest predictions from the MALARIA_RISK_FORECAST table. 3. **System:** Merges spatial data with forecast scores and applies color-coding logic (e.g., Red for > 0.80). 4. **User:** Views the fully rendered, interactive map with click-able zones.
Post-condition	• The user successfully views the current spatial risk layout. No database state changes occur (Read-Only).

Create User_Risk_Alert

This use case involves the automated system generating warnings when the AI predicts an outbreak. The main success scenario includes the system detecting a high-risk forecast, identifying the users responsible for that zone, and creating an active alert record.

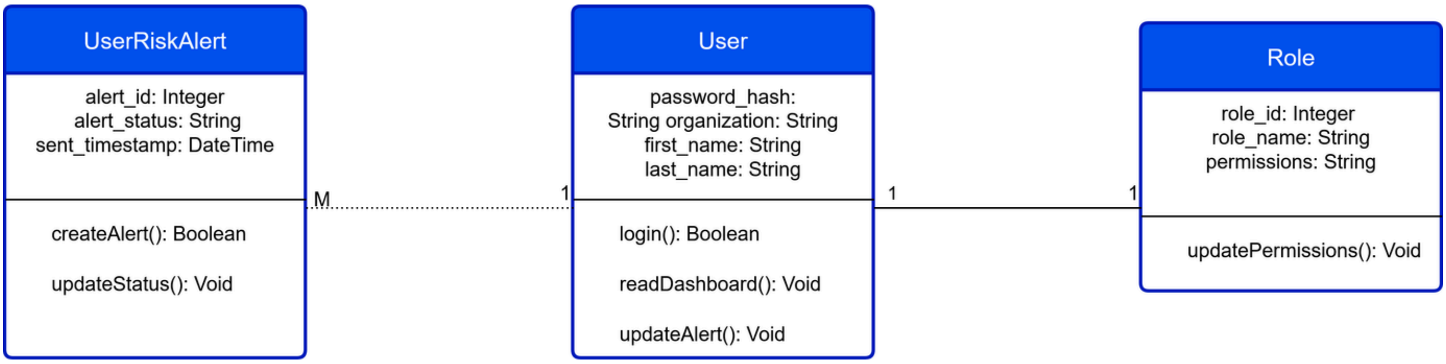
Use Case Name	Create User_Risk_Alert
Triggering Event	The ML Pipeline saves a forecast with an outbreak_probability_score exceeding the minimum threshold.
Actor(s)	
Related Use Cases	• Create Malaria_Risk_Forecast • Update Alert Status
Pre-condition(s)	• Forecasts exist for the current date range.
Flow of Activities	1. **ML Pipeline:** Completes the Create Malaria_Risk_Forecast sequence with a high-risk score. 2. **System:** Intercepts the high-risk trigger and initiates the alert workflow. 3. **System:** Queries the database for all users assigned to the affected zone_id. 4. **Database:** Retrieves user_id and email_address from the USER table. 5. **System:** Formats the alert payload (status = 'Active') and sends an email notification. 6. **System:** Executes the INSERT command to track the alert state. 7. **Database:** New records are added to the USER_RISK_ALERT table.
Post-condition	• New active alert records are created in the USER_RISK_ALERT table linking the user_id to the forecast_id, with the status set to 'Active'.

DOMAIN MODEL CLASS DIAGRAM

The following Domain Model Class Diagram illustrates the conceptual structure of the platform. It maps out the core business entities, such as users, spatial zones, environmental observations, and predictive forecasts, along with their key attributes and relationships. This visual blueprint abstracts the technical database details to provide a clear understanding of the underlying business logic, ensuring all system components align with the overarching goals of spatial health analytics and proactive risk alerting.

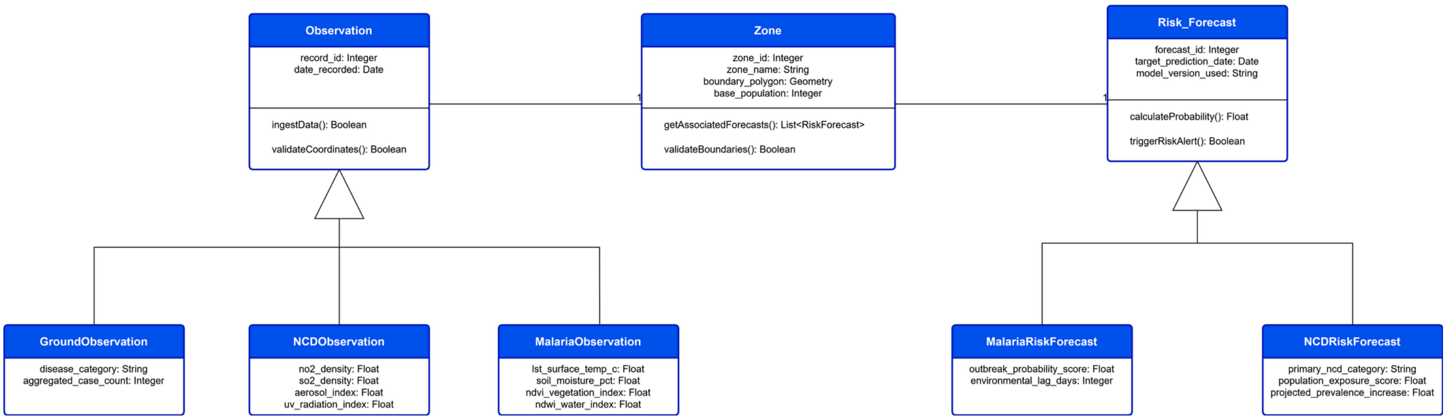
User Management DMD

The following Domain Model Class Diagram illustrates the conceptual structure of the client-facing frontend system. It maps out the core business entities responsible for access control, user engagement, and notifications, specifically focusing on users, roles, and risk alerts. This visual blueprint abstracts technical implementation details to highlight how user interactions and state changes are managed within the application layer.



Observation Prediction ERD

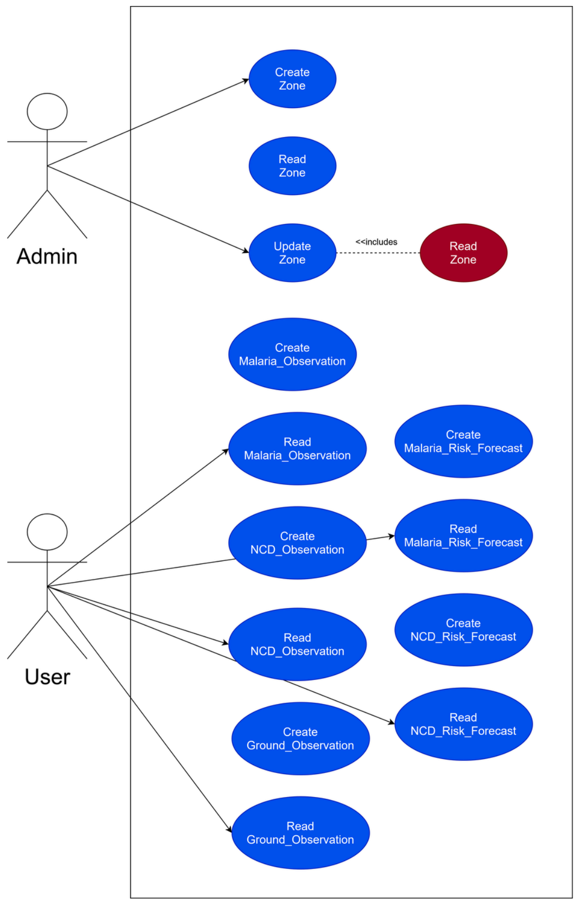
The following Domain Model Class Diagram outlines the conceptual structure of the core backend engine. It defines the primary business entities that drive the platform's analytical capabilities, including spatial zones, environmental and clinical observations, and machine learning risk forecasts. By utilizing object oriented principles such as inheritance for the various observation and forecast types, this diagram provides a clear blueprint of the underlying business logic and data relationships that power the predictive pipelines.



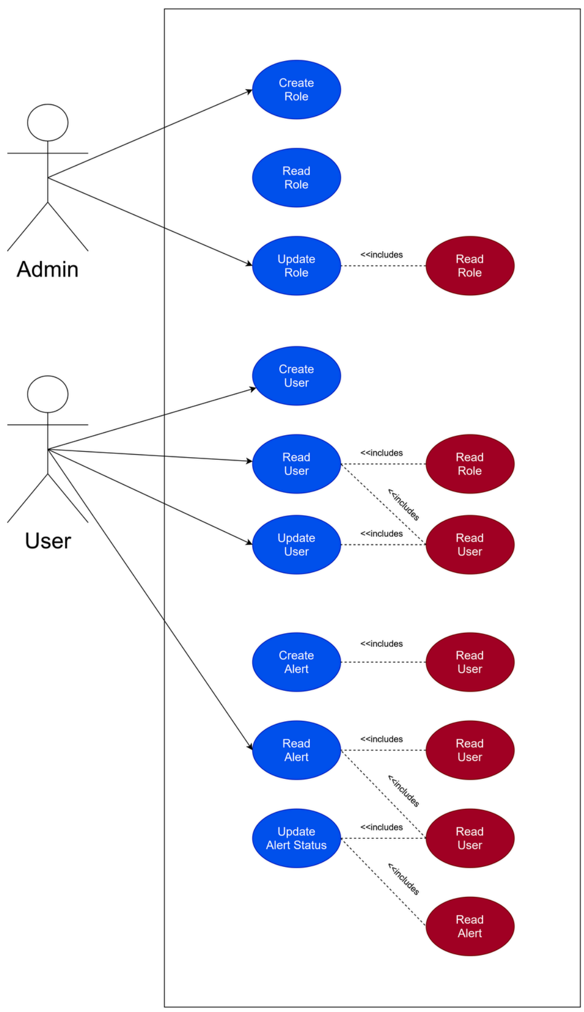
USE CASE DIAGRAM

The following Use Case Diagram provides a visual overview of the functional requirements of the platform. It illustrates the interactions between the defined system actors and the core processes they perform. This diagram offers a clear picture of the platform capabilities, such as viewing spatial risk maps, generating predictive forecasts, logging health data, and managing user access, ensuring all stakeholders easily understand the scope and boundaries of the system.

Observation Prediction Use Case Diagram



User Management Use Case Diagram



SUPPLEMENTARY SPECIFICATION

INTRODUCTION

This document outlines the supplementary specifications for the Earth Observational Health Analytics (EOHA) system. It details the structural design of the system's reporting modules, provides visual evidence of report implementations, and defines the supplementary business and domain rules governing Earth Observation (EO) data processing and health informatics.

REPORTING

Report Designs and Motivation

The reporting module is a critical component for public health stakeholders, translating complex EO data and Machine Learning (ML) outputs into actionable insights.

Report Name	Design Overview	
Malaria Risk Assessment	Centralized geospatial heatmap of Africa, supported by real-time alert logs and risk factor breakdowns.	Enables early detection of malaria outbreak-prone areas by correlating surface temperature and soil moisture.
NCD Trajectory Forecast	Trendline graphs tracking NO2 vs PM2.5, cardiovascular risk metrics, and a '5th Wave' monitoring index.	Assists policymakers in identifying long-term respiratory and cardiovascular risk trends driven by urban growth.
Environmental Factors	Time-series visualizations comparing seasonal variables (rainfall, humidity) with spatial anomaly detection.	Isolates the specific environmental drivers responsible for anomalous disease vector breeding.
Model Performance Audit	Statistical breakdown featuring Accuracy, Precision, Recall matrices, and ROC curves.	Ensures continuous transparency, model validation, and trust in the Ecological Niche Modelling algorithms.

NARRATIVE DESCRIPTION OF REPORTS

Malaria Risk Assessment Dashboard

This primary diagnostic tool visualizes the geographic probability of malaria transmission. The interface is anchored by a real-time geospatial heatmap of the African continent. Flanking the map are modular components including an "Active Monitoring" statistical summary, a breakdown of contributing risk factors (such as standing water vs. surface temperature), and a real-time alert log. These visual outputs empower ministries of health to allocate resources preemptively rather than reactively based on machine-flagged hotspots.

NCD Trajectory Forecast Dashboard

Designed to tackle the high mortality rate attributed to non-communicable diseases, this report correlates environmental pollution metrics with epidemiological forecasts. The dashboard prominently displays comparative line graphs for NO2 vs. PM2.5 exposure, alongside cardiovascular risk projections. Furthermore, it incorporates the proprietary "5th Wave Index" to track overlapping environmental and health crises, indicating exactly where NCD burdens are likely to spike alongside rapid urbanization.

Environmental Factors Analysis Dashboard

A purely scientific reporting module that displays the raw and processed variables feeding the ML engine. It tracks key metrics like temperature, rainfall, and humidity against historical baselines using interactive time-series charts. The interface also includes a spatial mapping feature specifically configured to highlight isolated environmental anomalies, utilizing Fourier processing to filter out standard seasonal noise from genuine ecological shifts.

Model Performance Audit Dashboard

An administrative report utilized by the Master's and PhD data science team. It evaluates the predictive reliability of the Principal Component Analysis (PCA) and niche models. The interface visualizes core ML metrics—including dynamic dials for Precision, Recall, and overall Accuracy—alongside Receiver Operating Characteristic (ROC) curves. It is designed to flag ”model drift” instantly if the algorithmic accuracy drops below the accepted 90% threshold.

BUSINESS RULES

The following business rules establish the operational, compliance, and quality assurance frameworks required for the platform to function securely and effectively. These directives govern the system's overarching behavior rather than its underlying data structure. Their primary purpose is to ensure regulatory compliance (specifically POPIA), maintain the timeliness and integrity of ingested datasets, enforce strict automated quality controls, and mitigate clinical risk by clearly defining the boundaries of the platform's predictive intelligence

01

BR-01

Risk predictions must be systematically refreshed at scheduled intervals (minimum 24-hour cycles) using the latest SANSA and Google Earth Engine API pulls.

02

BR-02

The platform must display a persistent disclaimer stating that outputs are predictive intelligence and must be validated before enacting clinical medical interventions.

03

BR-03

All user access, data ingestion, and model retraining activities must be immutably logged to ensure POPIA compliance and full auditability.

04

BR-04

No external datasets may be ingested without passing the automated ”Quality Test” protocol defined in Milestone 4

DOMAIN RULES

The following domain rules define the epidemiological, geospatial, and mathematical constraints that govern the platform's core algorithmic logic. Unlike general business rules, these directives are strictly tied to the scientific and subject-matter expertise required for public health modeling. Their primary purpose is to ensure that all data preprocessing pipelines, machine learning predictions, and statistical outputs adhere rigidly to established scientific principles specifically regarding disease transmission mechanics, environmental exposure modeling, and the accurate presentation of statistical confidence.

01

DR-01

Malaria pathogen transmission risk is mathematically constrained by specific environmental thresholds (e.g., specific combinations of standing water proximity and elevated surface temperatures).

02

DR-02

Non-communicable disease risk forecasting must weigh long-term environmental exposure metrics heavier than short-term anomalies..

03

DR-03

All Machine Learning predictions displayed on the user interface must be accompanied by a calculated probability score/confidence interval

04

DR-04

Environmental raster data and demographic vector data must be perfectly spatially and temporally aligned before entering the Principal Component Analysis pipeline.

DATA ARCHITECTURE

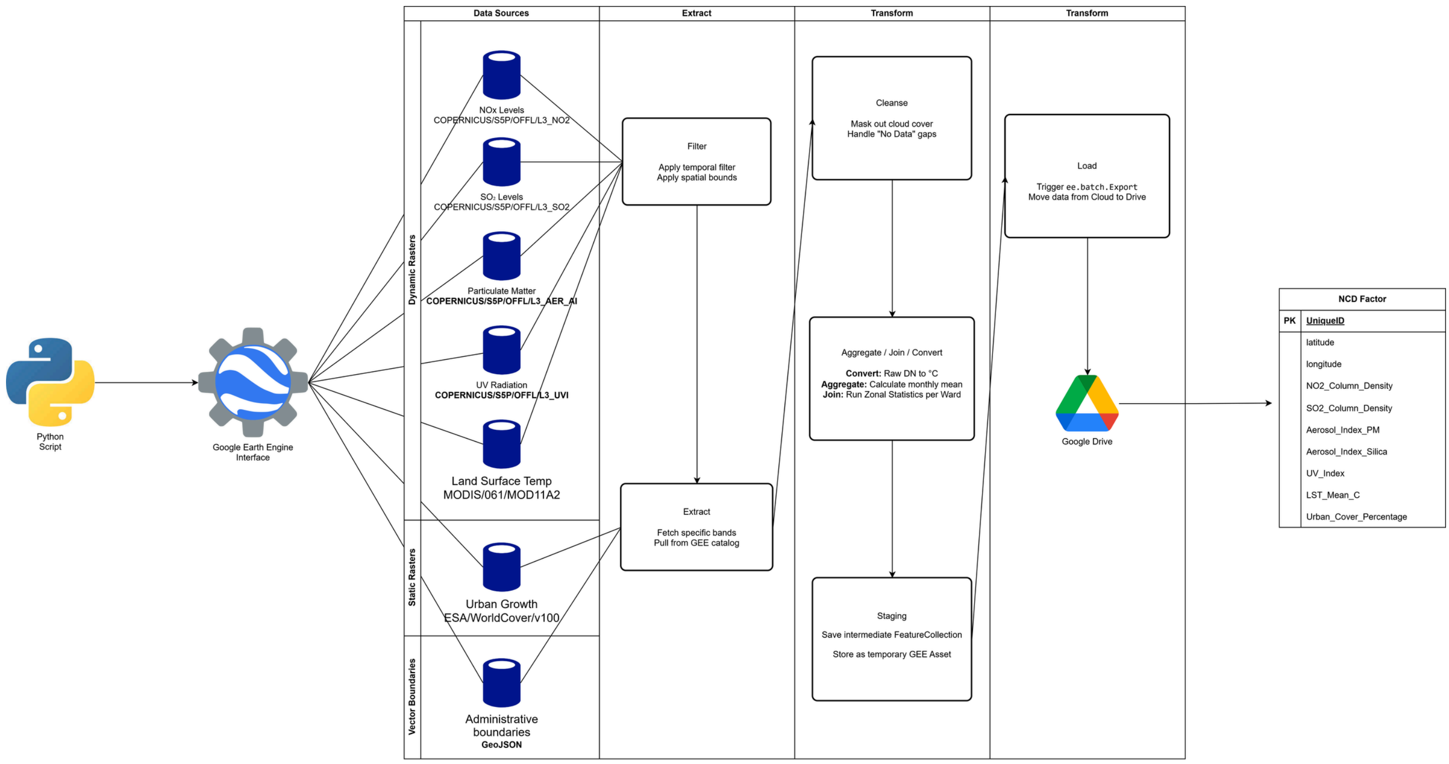
OVERVIEW

This document outlines the system architecture for a data-intensive, real-time data mining platform designed to act as machine surveillance for public health. The platform leverages Earth Observation (EO) data to address two of the most significant health threats in Sub-Saharan Africa: Malaria and Non-Communicable Diseases (NCDs). Currently, Sub-Saharan Africa accounts for approximately 95% of global malaria deaths, while NCDs rapidly contribute to over 37% of all deaths in the region. By tracking environmental risk factors and water-borne pathogen transmission in real-time, this platform aims to inform governmental public health policies and interventions. Processing is orchestrated using Python and the Google Earth Engine (GEE) API, supported by the high-performance computing capabilities of the CSIR. The platform is designed to be highly customizable, allowing specific districts to plan surveillance programs tailored to their unique geographic needs.

PIPELINE ARCHITECTURE DIAGRAM

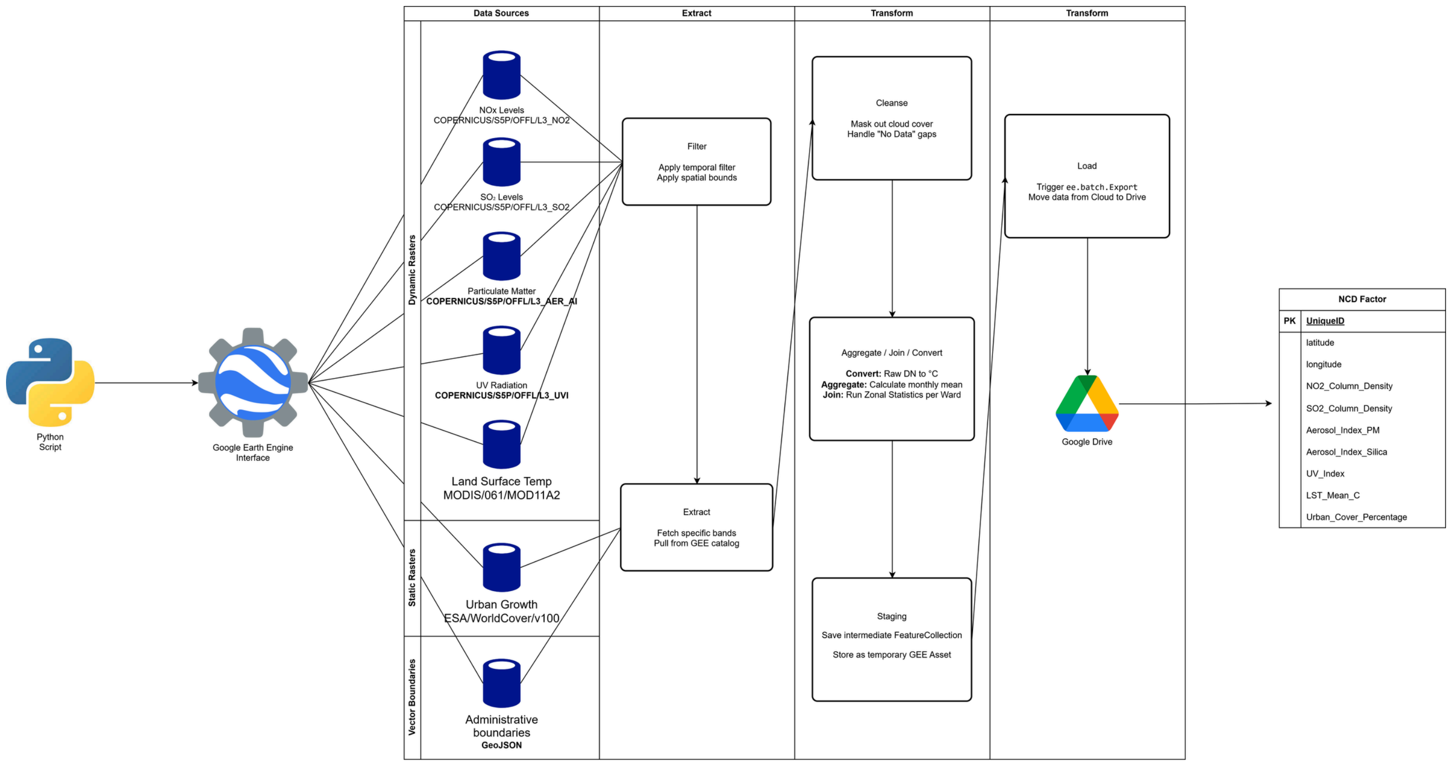
Non-Communicable Disease (NCD) Environmental Risk Data Pipeline

This diagram details the Extract, Transform, and Load (ETL) architecture specifically designed for communicable disease tracking, focusing on Malaria. It illustrates how a Python script triggers the Google Earth Engine API to aggregate critical epidemiological indicators. The pipeline ingests dynamic environmental rasters (such as surface temperature and soil moisture) alongside static demographic and habitat data. After applying spatial and temporal filters, cloud masking, and zonal statistics, the data is batched into Google Drive to populate the final Malaria Factor schema for downstream predictive analysis.



Malaria & Water-Borne Pathogen Environmental Risk Data Pipeline

This diagram maps the ETL pipeline tailored for Non-Communicable Disease (NCD) risk assessment. While structurally similar to the communicable disease workflow, this pipeline ingests a distinctly different set of atmospheric and environmental data sources crucial for NCD modeling. It extracts dynamic raster data from Copernicus/Sentinel satellites specifically tracking air quality metrics like NO2, SO2, particulate matter, and UV radiation along with land surface temperature and urban growth boundaries. The transformed output is compiled into the NCD Factor schema to analyze the impact of long-term environmental exposures on public health.



SYSTEM PROPERTIES & COMMERCIAL VALUE

Real-Time Surveillance

The platform acts as an interactive machine surveillance system, providing real-time data to ministries of health to monitor changing disease patterns.

Customized District-Level Reporting

The system rejects a blanket approach; analytics are refined so that individual districts can uniquely plan their surveillance programs.

Commercial & Academic Viability

The output data supports commercial API integration, intergovernmental agency platforms, and drives postgraduate research through partnerships.

OUTPUT FEATURE TABLES

NCD Environmental Indicators

The following data dictionary outlines the atmospheric and environmental variables extracted to model Non-Communicable Disease (NCD) risk. These indicators specifically target long-term environmental exposure metrics such as air pollution (NO2, SO2, particulate matter) and urban heat islands which are critical for forecasting respiratory, cardiovascular, and other environmentally aggravated health conditions.

Field	Roles
UniqueID	District/Ward identifier
latitude	Location coordinate
longitude	Location coordinate
NO2 Column Density	Average NO2 exposure
SO2 Column Density	Average SO2 exposure
Aerosol Index PM	Particulate proxy
UV Index	UV radiation level
LST Mean C	Land surface temperature
Urban Cover Percentage	Urban land cover ratio

Malaria Environmental Indicators

This data dictionary details the geospatial and climatic variables used to calculate Malaria transmission risk. The selected indicators are explicitly chosen to map the ecological conditions necessary for vector-borne diseases, focusing on the specific combinations of surface temperature, soil moisture, vegetation, and standing water that support mosquito breeding habitats and human-vector interaction.

Field	Roles
UniqueID	District/Ward identifier
latitude	Location coordinate
longitude	Location coordinate
LST Surface C	Surface temperature
Soil Moisture	Soil moisture level
NDWI water	Water index
Habitat Vegetation Index	Vegetation suitability
Population Density Per KM2	Population density